YASS: enhancing the sensitivity of DNA similarity search

Laurent Noé* and Gregory Kucherov

LORIA/INRIA-Lorraine, 615 rue du Jardin Botanique, 54602 Villers-les-Nancy, France

Received March 23, 2005; Accepted April 15, 2005

ABSTRACT

YASS is a DNA local alignment tool based on an efficient and sensitive filtering algorithm. It applies transition-constrained seeds to specify the most probable conserved motifs between homologous sequences, combined with a flexible hit criterion used to identify groups of seeds that are likely to exhibit significant alignments. A web interface (http://www.loria.fr/projects/YASS/) is available to upload input sequences in fasta format, query the program and visualize the results obtained in several forms (dot-plot, tabular output and others). A standalone version is available for download from the web page.

INTRODUCTION

Modern bioinformatics relies heavily on alignment programs and motif discovery tools, and numerous comparative genomics projects need ever more precise and faster tools for comparing two or several genomic sequences with different resolutions.

Except for small sequences, the exact local alignment algorithm of Smith and Waterman (1) is not frequently used, and most alignments are obtained using heuristic alignment tools such as FASTA (2), FLASH (3), BLAST (4,5), BLASTZ (6) and PatternHunter (7,8). All these methods introduce a tradeoff between two competing parameters: *selectivity* (or *specificity*) directly affecting the speed of the algorithm and *sensitivity* affecting its precision (i.e. the number of relevant alignments missed). Achieving a good trade-off between sensitivity and selectivity is the key issue in local alignment tools. The recently introduced *spaced seeds* technique (7,8) allows an increase in sensitivity without loss in selectivity. This innovation triggered various studies (9–15) related to the usage, design and generalizations of spaced seeds.

In this note, we present YASS (Yet Another Similarity Searcher)—a new software for computing local alignments

of two DNA sequences—and its web server (http://yass. loria.fr/interface.php). Compared with other tools, YASS is based on two innovations. The first is a new spaced seed model called *transition-constrained seeds* that takes advantage of statistical properties of real genomic sequences. The second feature is a new statistically founded *hit criterion* that controls the formation of groups of closely located seeds that are likely to belong to the same alignment. An implementation of these improvements, reported here, provides a fast and sensitive tool for local alignment of large genomic sequences.

DESCRIPTION

Web interface

The main user input (Figure 1A) consists of one or two sequences in fasta format either chosen from a predefined database or uploaded to the web server.

Once sequences have been selected, the user can run the program right away with all other parameters set by default. Alternatively, the user can set other parameters such as the scoring matrix or gap penalties (preselected matrices are proposed), and specify the DNA strain to be processed (direct, complementary or both). The user can also choose to display complete alignments rather than only alignment positions.

More advanced parameters are available for expert users. For example, the right choice of the seed pattern can increase the search sensitivity considerably provided that some knowledge of target alignments is available (10–14). The web interface provides a preselection of seeds including three transition-constrained seeds, one providing a good performance compromise between coding and non-coding sequences, and the other two tuned respectively for non-coding and coding regions. The accompanying Hedera program (http://www.loria.fr/projects/YASS/hedera.html) is also provided for advanced users in order to design new seed patterns according to different probabilistic models of alignments (15).

Finally, the user can specify some statistical parameters of target alignments, such as the assumed substitution rate or indel rate. These parameters control the hit criterion, i.e. the rules for grouping together closely located seeds to detect similarities.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

^{*}To whom correspondence should be addressed. Tel: +33 3 83 59 30 11; Fax: +33 3 83 27 83 19; Email: Laurent.Noe@loria.fr

[©] The Author 2005. Published by Oxford University Press. All rights reserved.

Please enter DNA sequence(Enter one or two files, or selec								
Enter one or two files, or selec-		5-195-5-00 MM (\$-2079) Ap. 20-20 (60						
selected), then it is compared	t one or two available sequences. If only or with itself.	ne sequence is submitted (or						
Select file(s) to be uploaded :								
file 1:	Parcourir	add sequence (n						
file 2 :	Parcourir	add sequence (n						
database:choose spec	cleschoose sequence	add sequence						
	<u></u>	(*) : uploaded data is limited to 3 Mb overall	14253 A0660A	60 425350 ABBAAATTTCAGA	425340 4253 ATCCOTCATAGCOCCA	330 425320 YTCCATTGGAAAGATG	425310 486AGCCAG	425300 42529 COTAGGATACAGGACGA
		Market and a second and a second	ACAGCA	AAABACTTTGCAGA 0080 1230090	AAAAGTTATTGCACCA	AANTWTTGAAAGAATG	0AAAAGAC 1230130	AGACGCATACCAAAAGA 1230140 1230
Selected Sequence(s):			(4252 075CA0	80 425270 CCADATTOCCAAAGA	1425260 14252	250 425240 PRACCOTTOCCAGAGAA	425230 77400000	[425220 42521
1000 1000 1000 1000		remove sequence	TTTCGA	TOGOTTAGCAAAA	TOGOTTTTTTGGGCAT	rescrattccteaaaa	TACCOTCC	CATGGGCGCAGATGCCC
		remove sequence						1230220 1230 425140 42513 TOTTTOTGGTTOAGAAG
			00A000 - - CAACAG	GANTAGTCGCAGAG : : : FINTTGTCGCAGAG	CAGATAGCAAGGAAGG - : - -	BATCCGACCGCATCCA 	TACCOGTGC	TOTPTOTOGTTGACAAC TOTPTOTAGTTGACAAC
			(123 (4251	0240 1230250	1230260 123	30270 1230280 390 425080	11230290	1230300 1230 425060 42505
	Parameters		TGGTCG	INTETENTEGEGAN	GIACOGAAATGAGAAG	STOTICSGASATATO	TTAGGAAG	GTTGCAAAGGGCGAATG
	Scoring system		1123	0320 1230330	11230340 123	30350 1230360	11230370	11230380 1230
coring matrix (match,trans	sition,transversion)	+3,-1,-4 •	14250 AAC990	40 425030 CATAGCCTCCACGG	425020 4250 AATCCGGACACGGTTC	010 425000 CABACOTTOCOGOCATI	424990 MACACCOT	424980 42497 TOCOARGAGOTCCOGTO
Gap costs (opening,extensi	on)	12 💌 , 4 💌	AACTGG 123	CATAGGCTCTACAG 0400 1230410	ACTCCCCACATGGATG	CCATOTACCTCCCAT	GATACAGT 1230450	A0000TGAAGAAGGGCG 1230460 1230
E-value threshold		1E-D6 💌	14949	60 149495D	1494940 14949	930 1494990	494910	1494900 149499
	Display		IIII I AATATA		AAATCGTTCATAAGC	TTUTAAGGGATATCA	GGAAATCG	GCGGTGGCTTTCTTACT
DNA strain: process forward,	reverse complement or both reads of the fir	st sequence wd&rc	(123	0480 1230490	1230500 123	30510 1230520	11230530	11230540 [1230]
	nd statistics: display quick alignments along	g with some	GTGAAA	ACCAGGGAAAATGC	COSTICACGTGGTATE	ATCACTICTCTACATA	CCTATTCC	424820 42481 GACGATATCGAGATATC
statistics	315 CL 25 %	-	1123	0560 1230570	[1230580 [123	30590 1230600	11230610	11230620 1230
			14248 TCT00A	00 424790 88A8ATGGGCAGGG	424780 4247 A666TTCGTCAT0666	770 424760 PISTOCTOAGOTTCAA	424750 SATCTCCA	424740 42473 TOTTCCGGCAAAGTACC
	Advanced parameters (optional)		TOTOGA 123	86 40 12306 50	1230660 123	GCACACTAAGGTTTAK 30670 1230680	1230690	COTTCCAGCTAAATACT 1230700 1230
these opti	ons are provided for advanced use of the VASS t	ool (see <u>fg.acmg</u>)	14247 T00000	20 424710 AGGAGAACAGGGGC	424700 4246 TTCAACATAGTTCATC	90 424600 BAAGGTTTTGAATICG	424670 AAGGGGCC	424660 42465 TOMINICOGTONIATOD
aroun Size specify the group	Search parameters Size (the total number of single nucleotide	matches in	170GAC	- -	TICANTATAGTGCATG	3AAGGTTTTGAATTTG 3A750 1230760	CAGGGGCT 11230770	: : - TAMICICOGITATATCI 1230780 1230
he group of seeds); increasin	g group size speeds up the computation bu	t 14 • bp						
decreases sensitivity.								
decreases sensitivity. seed Pattern: specify the see	d Pattern ('#' stands for match, '@' stands for	or match or #@##-#-## •						
decreases sensitivity. seed Pattern: specify the see	d Pattern ('#' stands for match, '@' stands for stands for match, '@' stands for match,	pr match or #@##-#-## =						
decreases sensitivity. seed Pattern: specify the see transition , '-' stands for joker)	Sequence mutation parameters	pr match or #@##-#-### ▼ 10 ▼ %						
decreases sensitivity, seed Pattern: specify the see ransition , " stands for joker) indels: maximal indel rate in al	Sequence mutation parameters ignments	1,9,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,						
decreases sensitivity, seed Pattern: specify the see ransition , " stands for joker) indels: maximal indel rate in al	Sequence mutation parameters ignments	10 🗷 %						
seed Pattern: specify the see ransition , '-' stands for joker) ndels: maximal indel rate in al	Sequence mutation parameters ignments	10 9 %	A					
lecreases sensitivity, seed Pattern: specify the see ransition , " stands for joker) andels: maximal indel rate in al	Sequence mutation parameters ignments ate in alignments	10 9 %	A					
lecreases sensitivity, seed Pattern: specify the see ransition , " stands for joker) andels: maximal indel rate in al	Sequence mutation parameters ignments ate in alignments	10 9 %	А					
lecreases sensitivity, seed Pattern: specify the see seed Pattern: specify the see seed Pattern: stands for joker) noted to the seed of th	Sequence mutation parameters lignments are in alignments submit to YASS clear input	10 • % 25 • % Positions		onus di	Statis	tics		Mutations
decreases sensitivity. seed Pattern: specify the see shands for joker) to stands for joker in dels: maximal indel rate in al Mutation: maximal mutation re	Sequence mutation parameters ignments ate in alignments	10 ■ % 25 ■ % Positions ThemoplasmaThermoplasma_volcania (32125-36530)	um fas Size (31592/31383	orward/reverse niverse	Evalue Score 0 21230	9710.41 4510		Bias 3229, 3634, 2685
lecreaces sensitivity, seed Pattern: specify the see arransition , "stands for jokar) indels: maximal indel rate in al autation: maximal indel rate in al au	Sequence mutation parameters ignments to in elignments submit to YASS clear input	Positions ThemoplasmaThermoplasma_volcanin(1220078.1253/245).	um.fas Size 1 31592/31383 13817/13/71 10894/10618	reverse reverse	Evalue Score	9710.41 4510 6214.01 1800 7223.88 1120	1793 1020	3229, 3634, 2685 1609, 1121, 868 407, 847, 891
decreases sensitivity. seed Pattern: specify the see seransition, "stands for joker) indels: maximal indel rate in al Mutation: maximal mutation ra ### Them 1 (1823) 2 (2823) 3 (1926) 6 (1855) 6 (2026) 6 (1855)	Sequence mutation parameters ignments ate in alignments submit to YASS clear input clear input	Positions Themoplasma/Thermoplasma_volcanin (322.26.36302). (1512.56.3612). (1513.56113). (1514.266.1719.52). (15369.5313.4421.1. (457076.473.5314).	im.fas Size 31592/31383 1381/19217, 10894/10610 30991/2095/7 26415/26526 1252/317/465	reverse reverse forward reverse reverse	Evalue Score 0 21230 0 17958 0 15793 0 15490 0 15153 0 12965	9710.41 4510 6214.01 1803 7223.88 1123 7085.31 434 6931.19 3793 5930.54 1770	1793 1020 4673 4331 1018	3229, 3634, 2685 1609, 1121, 868 407, 847, 891 3079, 2384, 3559 2953, 1897, 3274 1144, 1029, 1415
ecreaces sensitivity. eed Pattern specify the see anistion, "stands for joker) ndels: maximal indel rate in al Autation: maximal mutation ra 1 (1322) 2 (4233) 3 (13324) 6 (1355) 6 (1355) 8 (1349) 8 (1349)	Sequence mutation parameters ignments ate in alignments submit to YASS clear input clear input	Positions ThemoplasmaThermoplasma_volcaning	31592/31383 33592/31383 3351/13/217, 10884/10617 20991/309627 26415/263/26 21392/20329 3306/3399	reverse reverse forward reverse reverse forward forward	Evalue Score 0 21230 0 17958 0 15793 0 15490 0 15153 0 12965 0 12466 0 12150	9710.41 4510 9710.41 4510 8214.01 1880 7223.98 1122 7085.31 4340 6931.19 379 5930.54 1770 5702.33 2720 5557.81 1277	1793 1020 4673 4331 1018 3202 1301	Blas 3229, 3634, 2685 1609, 1121, 668 407, 647, 691 3079, 2384, 3559 2953, 1897, 3274 1144, 1029, 1415 2174, 1926, 1822 963, 840, 755
lectreases sensitivity, seed Pattern specify the see analysis on, "stands for jokar) andels: maximal indel rate in al full seed pattern maximal mutation residual mutation res	Sequence mutation parameters ignments to in alignments submit to YASS clear input submit to YASS clear input clear inpu	Positions ThemoplasmaThermoplasma_volcanis (32126-3639) (14208-1243) (14126-171952) (14508-18-171952)	3152(21138) 3152(21138) 3181(7)3717 10004/10610 32981-20957 26415(2526) 21302-20329 3006(7039) 10117/10180 1249917(005)	reverse reverse reverse forward reverse forward forward forward forward	Evalue Score 0 21230 0 17958 0 15793 0 15490 0 15153 0 12965 0 12160 0 12150 0 10791	9710.41 4511 6214.01 4511 7723.08 1122 7065.31 434 6931.19 373 5930.54 1777 5702.33 2720 5557.61 1277 5403.59 1423 4936.30 1766	5 1793 5 1020 6 1020 9 4673 1 4331 1 1010 0 3202 7 1301 1 545 7 2004	3229, 3634, 2685 1609, 1121, 868 407, 847, 891 3079, 2384, 3559 2953, 1097, 3274 1144, 1029, 1415 2174, 1926, 1822 983, 840, 755 1451, 703, 616
decreases sensitivity. decreases sensitivity. decreases sensitivity. decreases sensitivity. decreases sensitivity. decreases sensitivity the see a ransition. decreases sensitivity. decrea	Sequence mutation parameters ignments to in elignments submit to YASS clear input oplasma/Thermoplasma_acidophilum.fas oplasma/Thermoplasma_acidophilum.fas opl.12(2)(2)(3)(3)(3)(4)(4)(4)(4)(4)(4)(4)(4)(4)(4)(4)(4)(4)	Positions ThemoplasmaThermoplasma_volcanis (322125-365397). (1320078-1243794). (65198-39813). (1320078-1243794). (13208-1136821). (13208-1136821). (13208-1136821). (13208-1136821).	31532/31383 13817/32/17 10864/10610 30981/30957 25415/25526 12522/072455 2306/3309 10117/10100	reverse reverse forward reverse forward forward forward forward	Evalue Score 0 21230 0 17950 0 15793 0 15490 0 15153 0 12965 0 12466 0 12150 0 11013	9710.41 4510 9710.41 4510 0214.01 1806 7223.88 1122 7065.31 4345 6931.19 3793 5930.54 1770 5702.33 2725 5557.81 1277 5403.59 1423	5 1793 5 1020 9 4673 1 4331 1 1010 0 3202 7 1301 5 1545 7 2004 2 1443	Bias 3229, 3634, 2685 1609, 1121, 868 407, 847, 891 3079, 2384, 3559 2953, 1897, 3274 1144, 1029, 1415 2174, 1926, 1822 983, 840, 755 1451, 703, 816
tecreaces sensitivity. tece Pattern specify the see annistion , ** stands for joker) ndels: maximal indel rate in al Autation: maximal	Sequence mutation parameters ignments to in alignments submit to YASS clear input	Positions ThemoplasmaThermoplasma_volcanis (32125-93597) (1059098-1385921) (66136-36813) (1159098-1385921)	um.fas Sice 1 13817.211832. 13817.211832. 13817.2117. 108947.0616. 20991.0696. 2182.0703.59. 2182.0703.59. 1282.0703.59. 11915.0703.28. 1282.0703.59. 11915.0703.28. 1282.0703.59. 17915.0703.28.	reverse reverse forward reverse forward forward forward forward reverse reverse forward	Evalue Score 0 21230 0 17958 0 15793 0 15490 0 15153 0 12965 0 12466 0 12150 0 11013 0 10791 0 10394	## Score ## 9710.41 4510.01 1800 77221.00 1122 77085.31 4344 6931.19 378; 5790.54 175.76 1277 5403.59 1422 4736.30 176.4754.73 1333 4721.81 2310.00 1003	5 1793 5 1020 9 4673 1 4331 1 1010 0 3202 7 1301 5 1545 7 2004 2 1443	329, 3634, 2685 1699, 1121, 668 407, 647, 891 3079, 2364, 3559 2933, 1897, 3274 1144, 1029, 1415 2174, 1926, 1822 983, 840, 755 1451, 703, 616 1441, 1430, 840 986, 576, 1211
ecreaces sensitivity. eed Pattern specify the see anistion, "stands for joker) ndels: maximal indel rate in al Autation: maximal mutation re 1 (13228 1 (13228 1 (10929 5 (2008) 6 (13558 7 (14952 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412) 1 (13412)	Sequence mutation parameters ignments to in alignments to in alignments submit to VASS clear input submit to VASS clear input submit to VASS	Positions ThemoplasmaThermoplasma_volcanis (252 \$\leftilde{x}\$ % ThemoplasmaThermoplasma_volcanis (252125-3532) (1230078-1231724) (1230078-1231724) (145086-171952) (145086-1	um.fas Sice 1359/2/1383. 1383/2/1383. 1383/2/1383. 1383/2/1383. 1389/10864/10810. 2099/10867. 24515/2525. 2189/20829. 1395/1089. 1395/1089. 1395/1089. 1595/1089. 1595/1089. 2595/1599. 2715/1089. 2715/1089.	neverse neverse neverse forward neverse forward forward forward forward forward forward forward reverse neverse forward reverse neverse forward neverse proverse forward neverse neverse forward forwa	Evalue Score 0	BIL Score #9710.41 4511 6021.401 1000 621.40	5 1793 5 1020 6 1020 6 4673 1 4331 7 1010 7 1301 5 1545 7 2904 2 1443 2 2664 821 654 8 1417	Hiss 3229, 3634, 2685 1699, 1121, 968 407, 447, 691 3079, 2384, 3559 2953, 1897, 3274 1144, 1029, 1415 2174, 1926, 1822 983, 840, 755 1451, 703, 616 1441, 1499, 640 966, 576, 1211 1564, 2344, 1049 566, 330, 590 467, 198, 632
ecreases sensitivity. ed Pattern specify the see anistion, "stands for johar) ndels: maximal indel rate in al dutation: maximal mutation ra 1 (1323) 2 (24236 3 (13424 4 (1099) 5 (2000) 6 (1349) 1 (21427 1 (2	Sequence mutation parameters ignments to in elignments submit to YASS clear input submit to	Positions ThemoplasmaThermoplasma_volcanis (32125-3350) (1230978-1243724) (1230978-1243724) (141286-171952) (1143286-171952) (1143286-171952) (143286-171952) (143286-171952) (143286-171952) (143286-171952) (143286-171952) (143286-171952) (143286-171952) (143286-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952) (14328-171952)	imfas Stre 3158/2/1363 1381/2/1712 3091/10667 2481/2052 1252/17/46 2138/2039 2138/2039 1249/1205 1550/1569 1550/1569 6756/1749	reverse reverse forward reverse forward forward forward forward forward reverse reverse reverse forward reverse reverse reverse forward reverse	Evalue Score 0 21593 0 17950 0 17950 0 15793 0 15793 0 15183 0 12865 0 12466 0 12150 0 11013 0 10394 0 10394 0 10394 0 10394 0 10394 0 10394 0 10394 0 7953	Bit Score # 9710.41 4514 4514 4514 4514 4514 4514 4514 4	1793 1020 1020 1020 1020 1020 1301 1301 1545 2004 11443 12664 821 654 1417 993	Hiss 3229, 3634, 2685 1609, 1121, 060 407, 447, 931 3079, 2384, 3559 2835, 1037, 3274 1144, 1022, 3274 1274, 1326, 1822 933, 840, 755 1451, 703, 615 936, 576, 121 1504, 2341, 1049 636, 330, 590 467, 190, 632 500, 376, 420
ecreaces sensitivity. eed Pattern specify the see anistion , " stands for joker) ndels: maximal indel rate in al Autation: maximal indel rate in al 1 (3329) 2 (4238 2 (15324 4 (18099) 5 (2008) 6 (1388) 7 (1381) 1 (2192) 11 (2192) 12 (2215) 13 (2623) 14 (2723) 15 (2623) 16 (2633) 17 (2633) 18 (2633) 19 (2633) 19 (2633) 10 (2633) 11 (2633) 12 (2633) 13 (2633) 14 (2633) 15 (2633) 16 (2633) 17 (2633) 18 (2633) 19 (Sequence mutation parameters ignments to in alignments submit to YASS clear input poplasma/Thermoplasma_acidophilum.fas 2,115-61, 2,115-61, 2,115-61, 3,15	Positions ThemoplasmaThermoplasma_volcania(382125-285507). (1250725-185507). (06136-38613). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550765-185507). (1550767	um.fas Size 138/17/37/17/ 138/17/37/17/ 138/47/37/17/ 138/47/37/17/ 138/47/37/ 128/47/37/ 138/47/ 138/4	reverse reverse reverse forward reverse forward forward forward forward reverse revers	Evalue Score 0 21250 0 17950 0 17950 0 15793 0 15793 0 15183 0 12865 0 12466 0 12150 0 11013 0 10791 0 10392 0 10200 0 6530 0 7953 0 7955 0 7745	Bit Score # 9710.41 4511 4512 4514 4514 4514 4514 4514 45	1793 1020 1020 1020 1020 1020 1301 1301 1545 2004 11443 12664 821 654 1417 993	Hiss 3229, 3634, 2695 1609, 1121, 668 407, 847, 891 3079, 2384, 3559 2933, 1897, 3274 1144, 1029, 1415 2174, 1926, 1822 983, 840, 755 1451, 703, 616 1441, 1490, 840 986, 576, 1211 1504, 2341, 10-49 696, 330, 590 467, 198, 632 964, 694, 825
ecreaces sensitivity. eed Pattern specify the see anistion, "stands for joker) ndels: maximal indel rate in al Autation: maximal mutation re 1 (3)229 2 (2)236 2 (2)236 3 (3)22 4 (10099 5 (2)200 6 (3)355 7 (1009) 10 (2)22 11 (2)23 13 (2)23 14 (2)23 15 (2)23 16 (2)23 17 (2)23 18 (2)23 19 (2)23 20	Sequence mutation parameters ignments to in alignments submit to VASS clear input submit to VASS clear input oplasma/Thermoplasma_acidophilum/fas opl.122.2091 2-811.56 2-	Positions ThemoplasmaThermoplasma_volcanin (323/125-265597). (1230078-1245794). (65198-39813). (114208-117952). (1158082-1384821). (456779-4793-261). (1158082-138492). (25682-314909).	um.fas Sire 138173712. 138173712. 1388410810. 3299110952. 281305352. 281305352. 281305352. 2130270322. 230673092. 1391570832. 1385015690. 6756372. 213127026. 2257712390. 55915590. 756670372. 12533127276.	reverse reverse forward reverse reverse forward forward forward forward reverse revers	Evalue Score 22230 0 17550 0 17550 0 15793 0 15490 0 15490 0 12965 0 12965 0 12965 0 10791 0 10324 0 10322 0 10200 0 8530 0 8540 0 7955 0 7746 0 7550 0 7565	Bit Score # 9710.41 4516 821.4.01 1602 7223.08 1002 7223.08 1002 7223.08 1002 7223.08 1002 7223.08 1002 7223.08 1002 723.08 10	1793 1020 1 4673 1 4331 1 1010 1 1010 1 3302 7 1301 1 1545 7 2004 1 1545 7 2004 1 2664 821 654 821 654 1 1890 995 1890 995 1909	Hiss 222, 3534, 2685 1609, 1121, 686 1609, 1121, 686 1407, 687, 691 3070, 2394, 3559 2353, 1897, 32274 1144, 1029, 1415 2174, 1926, 1822 933, 840, 755 1451, 703, 616 1204, 22341, 1049 566, 330, 559 167, 1981, 632 394, 825 394, 8
erceases sensitivity. et all attentions are sensitivity. et all attentions are sensitivity. et al. (1823) 1. (1823)	Sequence mutation parameters ignments to in alignments to in alignments submit to VASS clear input to VAS	Positions ThemoplasmaThermoplasma_volcanis (25.25.25.36) (1.230078-1245724) (1.230078-1245724) (1.230078-1245724) (1.230078-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.23008-1245724) (1.2308-1245724)	um.fas Sice 3 128179.2712. 128179.2712. 1089470.610. 3599170.954. 24815.65522. 2505.7999. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.85. 1282.072.05. 1282.0	reverse reverse reverse reverse reverse reverse reverse forward forward forward reverse revers	Evalue Score 22230 9 17250 1 17550 1 15793 9 15490 1 15490 1 15430 0 12565 0 12566 0 12166 0 10791 0 10394 0 10302 0 10200 0 6530 0 8540 0 7553 0 7746 0 7454 0 7454 0 6770	BH Score # 9710.41 4516 2714.01 100 1722.1.00 1122 7005.31 424 531.1.0 1722.50 1122 7005.31 424 531.1.0 1722 5557.81 1277 5507.81 1277 5403.69 178.30	1793 1020 14673 14331 1818 1202 1301 1545 2004 1443 12664 821 821 821 821 821 821 821 821 821 821	Hiss 222, 3534, 2695 1609, 1721, 668 407, 647, 691 3079, 2394, 3559 2593, 1697, 3274 1144, 1022, 1-115 174, 1926, 1822 963, 640, 755 1745, 1926, 1822 963, 640, 755 165, 1921 1564, 2341, 1049, 640 1566, 437, 753 762, 333, 685 1766, 427, 333, 263 2199, 2425, 1513 1464, 1701, 2433, 353 2199, 2425, 1513
ed Pattern specify the see earnston, "stands for joker) andels: maximal indel rate in al futation residual independent in all futation maximal independent in all futation residual residual in all futation residual re	Sequence mutation parameters ignments to in alignments to in alignments submit to VASS clear input clear inpu	Positions ThemoplasmaThermoplasma_volcanis (32125-335) (12309R-1231724) (12309R-1231724) (14728-175304) (14728-175304) (14308-171952) (14508-171952)	imfax Sirs 31502013892 31502013892 31502013892 315020120567 248150255 24815025 2482025 2482025 1181501882 128071285 1181501882 128071285 1280	reverse reverse forward forward forward reverse forward forward forward forward forward reverse reverse reverse reverse reverse reverse forward forwar	Evalue Score 21230 0 17950 0 17950 0 15490 0 15490 0 15490 0 15183 0 12965 0 12140 0 10181 0 10034 0 10034 0 10030 0 8530 0 8530 0 7748 0 77560 0 77464 0 7215 0 6770 0 6674	BH Score # 9710.41 4516 2714.01 100 2721.00 1122 7005.31 424 531.19 372 5330.54 1777 5702.53 272 5557.81 1277 5403.65 120 3351.65 643 3351	1793 14673 14673 14331 13202 1301 1301 1545 2004 21443 12664 821 654 1417 993 1890 995 1072 1999 1995 1072 1996 13245 1526 1526 1526 1526 1526 1526 1526 152	Hiss 222, 3544, 2685 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1675, 6874, 6815 1692, 2304, 3505 1625, 1307, 3274 1926, 1822 2734, 1926, 1822 2734, 1926, 1822 2734, 1926, 1822 2734, 1926, 1822 2734, 1926, 1822 2734, 1124, 1, 1490, 640 398, 576, 1271 1264, 2234, 1, 1049 268, 576, 1271 1264, 2234, 1, 1049 268, 477, 753 268, 477, 753 272, 333, 365 1156, 1533, 553 1156, 1533, 553 1156, 1533, 553 1156, 1533, 553 1156, 1533, 553 1156, 1533, 553 1156, 1533, 553 1157, 382, 1433 1077, 382, 1433 1077, 382, 1433 1077, 382, 1433 1077, 382, 1433
edea Pattern specify the see earnston, "stands for joker) and the see anston, "stands for joker) and the see anston, "stands for joker) and the see anston, "stands for joker) and the see anston of the see ansto	Sequence mutation parameters ignments to in alignments submit to YASS clear input	Positions ThemoplasmaThermoplasma_volcania(382128-285597). (128079-184794). (08198-28613). (108198-28613). (108198-28613). (108198-184794).	im.fas Stre 3158/2/1363 1381/2/371/L 3091/10967 24815/25/26 24815/26/26 24812/26/26 2180/2092 926/5/99 105/5/10/16/9 675/6/7/4 925/3/27/26 128/7/25/9 1115/90/2 128/7/25/9	reverse reverse forward reverse forward forward forward forward forward reverse revers	Evalue Score 21230 0 21230 0 17950 0 15793 0 15490 0 15490 0 15183 0 12866 0 12866 0 12965 0 12966 0 10731 0 10394	BH Score # 9710.41 4516 2721.40 1 1000 1722.30 1122 7005.31 434 5516 2722.30 1122 7005.31 434 5531.19 3722.30 1722 5557.31	1793 14073 14331 1331 13202 1301 1301 1301 1301 1301	Hiss 222, 3234, 2685 1699, 1121, 668 1699, 1121, 668 1699, 1121, 668 1699, 1221, 641, 691, 691, 691, 691, 691, 691, 691, 69
ecreases sensitivity. ed Pattern specify the see anistion, "stands for joker) indets maximal indet rate in al futation: maximal mutation re 1 03223 1 03223 2 04236 3 01322 4 10829 5 02088 6 01355 7 04382 11 0222 12 0235 13 02322 14 02322 15 02323 16 02323 17 02322 18 02322 19 02322 10 02322 11 02322 12 02332 13 02322 14 02322 15 02322 16 02322 17 02322 18 02322 19 02322 19 02322 10 02322 10 02322 11 02322 12 02322 13 02322 14 02322 15 02322 16 02322 17 02322 18	Sequence mutation parameters ignments to in alignments to in alignments submit to YASS clear input	Positions ThemoplasmaThermoplasma_volcania(332125-35597). (323125-35597). (06156-39813). (114226-1171952). (1250076-12537594). (06156-39813). (11560756-135914). (1356082-136913). (1356082-13691). (1356082-136909). (566082-63406). (145071-135909). (566082-63406). (145082-71-145902). (145082-71-14902-71).	um.fas Sire 1381732113833 1381732112 13884118613 23991199657 24984118613 2499121995 2499	Inverse inverse inverse inverse inverse inverse forward forward forward forward forward forward forward inverse invers	Evalue Score Sco	BII Score 2 9710.41 97	1 1793 1 1020 1 4673 1 4331 1 8100 1 3202 1 301 1 545 2 2004 821 654 1 1417 993 1 1890 995 1072 1 1909 1 1909	Hiss 222, 3544, 2685 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 619 1609, 112
### Times and the property of the see ansiton, "stands for joker) andels maximal indel rate in all full and	Sequence mutation parameters ignments to in alignments to in alignments submit to VASS clear input to VAS	Positions ThemoplasmaThermoplasma_volcanis (25272-583) ThemoplasmaThermoplasma_volcanis (252725-353) (1141286-1171952) (1143286-1171952) (1143286-1171952) (1143286-1171952) (1143286-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (114386-1171952) (119386-1171952)	um.fas Sice 1382.1383.12817.3717.2 10884-10810.2 2093.1298.1 2093.1298.1 2093.1 2095.1	Privation Privat	Evalue Score	BH Score 6 9710.41 971	1 1793 1 1020 1 4673 1 4331 1 1010 1 3202 1 1545 2 2004 2 1443 1 2564 821 1 654 1 1417 993 1 1890 995 1 1072 1 1909 1 3245 1 1909 1 1022 1 1022 1 1032 1 1032	Hiss 222, 3534, 2685 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 687 687, 687, 687, 687, 687, 687, 687,
ecreaces sensitivity. ecel Pattern specify the secenariston, "stands for joker) ndels: maximal indel rate in al alternation research and services	Sequence mutation parameters ignments to in elignments to in elignments to in elignments state in elignments clear input state in elignments clear input state in elignments clear input state in elignment	Positions ThemoplasmaThermoplasma_volcania(382128-38590). (382128-38590). (68198-38613). (68198-38613). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (158098-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590). (15819-138590).	im.fas Stre 3158/2/1158/3 1381/2/3/17/17 18091/1906/7 18091/1906/7 2451/2/2/3/5 2451/2/2/3/5 256/5/3904 1011/2/3/1809 12.8901/2/3/5 105/5/3/1/2/3/5 105/5/3/1/2/3/5 105/5/3/1/2/3/5 105/5/3/1/2/3/5 105/5/3/1/2/3/5 105/5/3/1/2/3/5 105/5/3/2/3/5 105/5/3/2/3/5 105/5/3/5/3/5 105/5/3/5/5/5 105/5/3/5/5/5/5 105/5/5/5/5/5/5/5/5/5/5/5/5/5/5/5/5/5/5/	Inverse inverse inverse inverse inverse inverse forward covered forward forward forward inverse invers	Evalue Score	Bit Score # 1 100 10	1 1793 1 1020 1 4673 1 4331 1 10110 1 3202 1 3202 1 1301 1 1545 2 1264 821 1 2264 821 1 2443 1 2264 821 1 1493 995 1 1890 995 1 1909 1 3245 1 200 1 1526 1 1613 1 1800 947 1 2111 1 1300 947 1 2111 1 1300 947	Hiss 222, 3634, 2695 [1609, 1121, 666 at 76, 761, 761, 761, 761, 761, 761, 761,
ecreaces sensitivity. eed Pattern specify the see anistion , " stands for joker) ndels: maximal indel rate in al Autation: maximal indel rate in al 1	Sequence mutation parameters ignments to in alignments site in alignments submit to YASS clear input submit to YASS submit to YASS submit to YASS submit submi	Positions Themoplasma Thermoplasma_volcania(32125-325391) (1258076-1257254), (06196-305391) (1558076-1257254), (06196-305391) (1568076-125084), (1568076-125086), (1568076-125086), (15680866), (15680866), (15680866), (15680866-1250866), (1568086666), (1568086666), (1568086666), (1568086666), (15680866666), (156808666666), (1568086666666666666666666666666666666666	mm.fas Stre 3158/271363. 138/17/3717. 1388/17/3717. 1388/17/3717. 1388/17/3717. 1388/17/3717. 1388/17/37/37/37/37/37/37/37/37/37/37/37/37/37	Inverse invers	Evalue Score	Bil Score 1970.11 1970	1 1793 1 1020 1 4673 1 4331 1 3202 1 301 1 3202 1 301 1 1545 2 2064 821 654 1 443 1 2664 821 1 447 993 1 1890 995 1 1909 1 3245 1 1613 1 2666 1 1613 1 1613 1 1616 1 1613 1 1616 1	Hiss 222, 3634, 2685 1609, 1121, 668 407, 647, 691 3079, 2344, 3595 2453, 1897, 3274, 619 1144, 1029, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929, 1415 2474, 1929
Lecreaces sensitivity. Lecephater specify the see anaston, "stands for joker) Indels: maximal indel rate in al Autation: maximal index in al Autation: maximal indel rate in al Autation: maximal index in al Autation: maximal indel rate in al Autation: maximal index in al Autation:	Sequence mutation parameters ignments to in alignments submit to YASS clear input submit to YASS clear input oplasma/Thermoplasma_acidophilum.fas opl.1262.091 2-8112.69 1-1.1562.1991 2-8112.69 1-826.1991 1	Positions Themoplasma_volcania(382128-385597). (382128-385597). (08138-39813). (1142280-11715821). (165138-39813). (1594758-1525984). (315685-22973). (315685-39813). (31578-3873). (31578-3873). (31578-3873). (31578-3873). (32578-3873).	mm.las Size 1381/13/21/2-1/2-1 1388/13/21/2-1 1388/13/21/2-1 1388/13/21/2-1 1388/13/21/2-1 1388/13/2-1	Inverse invers	Evalue Score Sco	BH Scott 100 1	1 1793 1 1020 1 4673 1 4531 1 1010 1 1010 1 2004 1 2004 1 2443 1 2564 1 1417 993 1 1890 995 1 1072 1 1890 995 1 2066 1 1613 1 1300 946 682 1 1613 1 1616 1 1616 692 1 1056 972	Hiss 222, 3544, 2695 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 619 1609, 112
# Them # Them indels: maximal indel rate in al and indels: maximal indel rate in al mutation: maxi	Sequence mutation parameters ignments to in alignments submit to VASS clear input	Positions ThemoplasmaThermoplasma_volcanin (382128-285890) (1820-2813) (114128-171982) (114128	imfas Stre 31582/31583 31582/	Inverse invers	Evalue Score	BH Scott Fig. Control Contro	5 1793 1 677 1 1020 1 1	Hiss 222, 3234, 2695 1609, 1121, 666 407, 847, 891 3079, 2394, 3595 2183, 1897, 3274, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1822, 1824, 1926, 1824
# Themselve to a seed Pattern specify the see transition, " stands for joker) indets maximal indet rate in al Mutation meedinal mutation re (1323) 1 (1323)	Sequence mutation parameters ignments to in alignments to in alignments steel in alignments steel in alignments steel in alignments submit to YASS steel input submit	Positions ThemoplasmaThermoplasma_volcania (382128-385302) (114128-1245324) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (158078-1253724) (169078-1253724) (169078-1253724) (169078-1253724) (169078-1253724) (199178-1253724) (199178-1263724)	im fas Stre 1382/21382, 1382/2	Inverse invers	Evalue Score Sco	BII Scott File Fi	5 1793 1 1020 1	Hiss 222, 3234, 2695 1609, 1121, 666 407, 647, 691 3079, 2394, 3595 1714, 1021, 1162 1715 1714, 1021, 1162 1715 1714, 1021, 1162 1715 1714, 1021, 1162 1715 1714, 1021, 1162 1715 1715 1715 1715 1715 1715 1715 171
# Themselve to a seed Pattern specify the see transition, " stands for joken" indets maximal indet rate in al Mutation: meedmal mutation re 1 (2329 5 (2423) 5 (2423) 5 (2423) 5 (2423) 5 (2423) 6 (2423) 6 (2423) 6 (2423) 7 (2423)	Sequence mutation parameters ignments to in alignments submit to VASS clear input	Positions ThemoplasmaThermoplasma_volcanis (25272-5828) (123076-1235724) (123076-1235724) (123076-1235724) (123076-1235724) (123076-1235724) (135086-13364621) (135086-13364621) (135086-1336421) (135086-1336421) (135086-1336421) (135086-1336421) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-134621) (135086-136621) (135086-1	mm.fas Sire 138173271352136932 13817327172 108844106103 2562136932 2562136932 25652139	Inverse invers	Evalue Score	BH Score 6710.41 5710.	5 1793 (1 1020) 1 1020	Hiss 222, 3534, 2685 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1121, 668 1609, 1209,

Figure 1. The input window (A) allows users to control most YASS parameters, from most basic to more advanced. Results are output in tabular format (B), with the possibility of displaying each sequence alignment (C).

Once the results are obtained, it is possible to generate a clickable dot-plot (Figure 2) where each alignment is linked to a URL with its text representation (Figure 1C). A tabular output (Figure 1B) is also available: alignments are sorted according to their E-value and linked to their text representation. Finally, the YASS output can also be downloaded in text format for further analysis.

Technical issues. The YASS server available at http:// yass.loria.fr/interface.php currently runs Apache 2.0.47

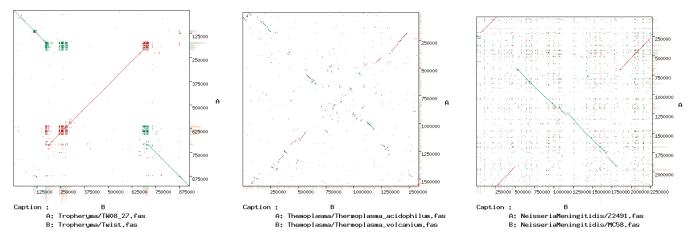


Figure 2. Three YASS dot-plots are shown, each obtained from pairs of closely related bacterial sequences. Green segments represent alignments of forward reads and red segments correspond to alignments between the reverse complement of one sequence and the forward read of the other.

(PHP and Perl-CGI modules) on a Linux Mandrake 9.2. Dot-plots are obtained with the GD graphical library interfaced to PHP. The YASS program has been developed in C and is distributed under the Gnu General Public License.

Owing to limitations of computational resources, some restrictions have been made on the web interface. For example, uploaded files are currently limited to 3 Mb, scoring systems can be chosen only among preselected ones and for each parameter a fixed range of possible values has been settled.

Standalone version

The standalone version is recommended for frequent users or those who need specific parameters to be set outside preselected values. It provides access to two other output formats, including a BLAST-like tabular format that can be used by existing postprocessing parsers. Note that YASS does not need one of the sequences to be preprocessed (formatdb command of BLAST), rather, it treats both sequences on the fly.

METHODS

Here we briefly outline the underlying principles of the YASS algorithm, including some novel features. For a more detailed presentation the reader is referred to (16) (http://www. biomedcentral.com/1471-2105/5/149/).

Seed model

Seeds are specified using a seed pattern built over a three-letter alphabet #, @ and -, where # stands for a nucleotide match, for a don't care symbol and @ for a match or a transition (mutation $A \leftrightarrow G$ or $C \leftrightarrow T$). The weight of a pattern is defined as the number of # plus half the number of @. The weight is the main characteristic of seed selectivity.

The advantage of transition-constrained seeds stems from the biological observation that transition mutations are relatively more frequent than transversions, in both coding and non-coding regions. Typically, biologically relevant alignments contain about the same number of transitions and transversions, whereas transitions are half as frequent in independently and identically distributed random sequences. Transition-constrained seeds increase the possible number of transitions in a hit relative to spaced seeds without the transition constraint, and this is done without loss of sensitivity or efficiency.

The sensitivity of a given seed has been estimated using the algorithm of (15), which is a generalization of the one proposed in (11). Two main alignment models have been considered: a Bernoulli model (13) assumed to simulate alignments of non-coding DNA and a hidden Markov model (10) assumed to simulate alignments of coding DNA. By default, YASS currently uses the seed #@# -- ## -- # - ##@# of weight 9, which provides a good compromise in detecting similarities in both coding and non-coding sequences. The standalone version of YASS allows users to specify their own seeds. Several preselected seeds are provided by the YASS web interface.

Hit criterion

YASS is based on a multi-seed hit criterion that defines a hit as a group of closely located and possibly overlapping seeds. Two seeds belong to the same group if they occur within a bounded distance or, on the other hand, are located at close dot-plot diagonals. Distance threshold parameters are computed according to probabilistic sequence models taking into account substitution and indel rates, similarly to models used in (17). Note that seeds of a group are allowed to overlap. An additional group size parameter sets a lower bound on the total number of individual matches and transitions of the group. Using the group size results in a flexible criterion that combines a guaranteed selectivity with a good sensitivity on both short and long similarities. More details on the hit criterion can be found in (16).

Comparative tests

To validate the better performance of transition-constrained seeds compared with ordinary spaced seeds, several comparative experiments have been presented in (16). Transitionconstrained seeds have been shown to be more sensitive with respect to some Bernoulli and hidden Markov models of alignments of coding and non-coding DNA [Tables 1 and 2 in (16)]. Moreover, transition-constrained seeds have been shown to be more sensitive in detecting alignments of real genomic sequences [Table 3 in (16)].

YASS has been compared with bl2seq (NCBI BLAST 2.2.6) according to several criteria: running time, number of significant alignments found (with E-value $\leq 10^{-6}$) and number of significant alignments found exclusively by one program and their total length [Table 4 in (16)]. The results show that YASS detects more significant alignments than bl2seq, within a smaller time for large DNA sequences.

CONCLUSIONS

In this paper, we have described YASS—a new DNA local alignment tool. The proposed web interface features several output formats suitable for a *coup d'oeil* analysis as well as for a deeper analysis of alignments. An upcoming release of YASS will include multi-seed indexing strategies and an optimized processor-cache algorithm.

ACKNOWLEDGEMENTS

The authors wish to thank Marie-Pierre Etienne and Pierre Valois for their help during early stages of this work, Mikhail Roytberg for his contribution to the seed design problem and Christophe Valmir for improving the YASS code. The authors have been supported by the ACI IMPBio of the French Ministry of Research. They thank the reviewers for their helpful remarks. Funding to pay the Open Access publication charges for this article was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences. J. Mol. Biol., 147, 195-197.

- 2. Lipman, D. and Pearson, W. (1988) Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA, 85, 2444-2448.
- 3. Califano, A. and Rigoutsos, I. (1993) Flash: a fast look-up algorithm for string homology. Proc. Int. Conf. Intell. Syst. Mol. Biol., 1, 56-64.
- 4. Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403-410.
- 5. Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., **25**, 3389-3402.
- 6. Schwartz, S., Kent, J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. Genome Res., 13, 103-107.
- 7. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics, 18, 440-445.
- 8. Burkhardt, S. and Kärkkäinen, J. (2001) Better filtering with gapped q-grams. In Proceedings of the 12th Symposium on Combinatorial Pattern Matching (CPM'01), 1-4 July, Jerusalem, Israel, LNCS 2089, Springer, pp. 73-85.
- 9. Li, M., Ma, B., Kisman, D. and Tromp, J. (2004) Pattern Hunter II: highly sensitive and fast homology search. J. Bioinform. Comput. Biol., 2,
- 10. Brejova, B., Brown, D. and Vinar, T. (2003) Optimal spaced seeds for hidden Markov models, with application to homologous coding regions. In Baeza-Yates, R., Chavez, E. and Crochemore, M. (eds), Proceedings of the 14th Symposium on Combinatorial Pattern Matching, 25-27 June, Morelia, Mexico, LNCS 2676, Springer, pp. 42-54.
- 11. Buhler, J., Keich, U. and Sun, Y. (2003) Designing seeds for similarity search in genomic DNA. In Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB'03), 10-13 April, Berlin, Germany, ACM Press, pp. 67-75.
- 12. Choi, K.P., Zeng, F. and Zhang, L. (2004) Good spaced seeds for homology search. Bioinformatics, 20, 1053-1059.
- 13. Keich, U., Li, M., Ma, B. and Tromp, J. (2004) On spaced seeds for similarity search. Discrete Appl. Math., 138, 253-263.
- 14. Kucherov, G., Noé, L. and Ponty, Y. (2004) Estimating seed sensitivity on homogeneous alignments. In Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE2004), May 19-21, Taichung, Taiwan, IEEE Computer Society Press, pp. 387–394.
- 15. Kucherov, G., Noé, L. and Roytberg, M. (2004) A unifying framework for seed sensitivity and its application to subset seeds. Rapport de recherche INRIA RR-5374. http://www.inria.fr/rrrt/rr-5374.html.
- 16. Noé, L. and Kucherov, G. (2004) Improved hit criteria for DNA local alignment. BMC Bioinformatics, 5, 149.
- 17. Benson, G. (1999) Tandem repeats finder: a program to analyse DNA sequences. Nucleic Acids Res., 27, 573-580.